OLAP4Tweets: Multidimensional Modeling of tweets

Maha Ben Kraiem, Jamel Feki, Kaïs Khrouf, Franck Ravat, Olivier Teste

OLAP4Tweets: Multidimensional Modeling of tweets

Maha Ben Kraiem[1,2], Jamel Feki[1], Kaîs Khrouf[1], Franck Ravat[2], Olivier Teste[2]

[1] MIR@CL Laboratory, University of Sfax
Airport Road Km 4, P.O. Box. 1088, 3018 Sfax, Tunisia
Maha.benkraiem@yahoo.com, Jamel.Feki@fsegs.rnu.tn,
Khrouf.Kais@isecs.rnu.tn.

[2] IRIT, University of Toulouse,
2, Rue du Doyen Gabriel Marty, 31042 Toulouse Cedex 9, France
{ Franck.Ravat, Olivier.Teste }@irit.fr

Abstract. Twitter, a popular microblogging platform, is at the epicenter of the social media explosion, with millions of users being able to create and publish short posts, referred to as tweets, in real time. The application of the OLAP (On-Line Analytical Processing) on large volumes of tweets is a challenge that would allow the extraction of information (especially knowledge) such as user behavior, new emerging issues, trends... In this paper, we pursue a goal of providing a generic multidimensional model dedicated to the OLAP of tweets. The proposed model reflects on some specifics such as recursive references between tweets and calculated attributes.
Keywords: Calculated attribute; Reflexive fact; OLAP model..

1 Introduction

Since its introduction in 2006, Twitter has evolved into an extremely popular social network and it has revolutionized the ways of interacting and exchanging information on the Internet. By making its public stream available through a set of APIs, Twitter has triggered a wave of research initiatives aimed at analysing and knowledge discovering from the data about its users and their messaging activities. We notice that the majority of works provided in the literature of this domain (analysis of tweets) are intended to answer specific tasks or needs. However, very few studies were interested in the multidimensional analysis of data from tweets so far. If it incorporates all the data issued from a tweet, this modeling could be a judicious opportunity to explore the tweets through an OLAP process. Tweets can be represented in a multidimensional way by extracting this metadata and social aspect of the messages. For this reason, we focus on the data warehouse as a tool for the storage and analysis of multidimensional and historized data. It thus becomes possible to manipulate a set of measures according to different dimensions which may be provided with one or more hierarchies. OLAP tools provide means to query and to analyze the warehoused information and produce reports at different levels of detail. These reports are computed using aggregate functions. Moreover, data from tweets have particular specificities,*e.g.*

inter-tweet relationships, calculated attributes. Hence, we consider them during the phase of multidimensional modeling.

The rest of this paper has the following structure. A state of the art of related works to tweets will be presented in Section 2. In Section 3, we propose an extension of our model [1]. In section 4, results and analyses for testing this multidimensional model on data extracted from tweets are presented.

2 Related Work

Twitter has largely contributed to the appearance of new issues related to the modeling and manipulation of data. In this context, the analysis of textual content of tweets and their meta-data is a promised research topic that has attracted the attention of an increasing community of researchers and has given birth to novel analysis areas, such as Social Media Analysis. A spectacular novel area of data analysis is that of the detection of tweets topics. An approach for disambiguating and categorizing the entities in the tweets aimed at discovering topics is described in [2]. The results obtained are used for determining users' topic profiles, and the possibility of analyzing them using OLAP techniques is not considered. The real-time identification of emerging topics in tweets is studied in [3]. Bursty keywords are extracted first, and then grouped to identify trends; however, trends are analyzed using a front-end with limited flexibility.

To the best of our knowledge, few researches have focused on the multidimensional modeling of tweets. Among these works, the one of [4] defined a multidimensional star model for analyzing a large number of tweets. However the proposed model was dedicated to a particular trend. In order to do this, the authors proposed an adapted measure, called "TF-IDF $_{adaptive}$", which identifies the most significant words according to levels of hierarchies of the cube (the location dimension). Nevertheless, their case study deals with a specific area: the evolution of diseases, by adding to their multidimensional model a dimension called 'MotMesh' (MeshWords).

A work sharing some similarities with ours is the one in [5] that presents architecture to extract tweets from Twitter and load them to a data warehouse. Conceptual models for Twitter streams from both OLTP and OLAP points of view are also proposed. However, both models are focused on the inter-relationships between tweets and between users, and little attention is paid to the social aspect of tweets (Tweet/tweet response). In [6] the authors model non-onto and non-strict topic hierarchies as DAGs (Directed Acyclic Graph) of topics. The proposed solution has higher expressivity with respect to traditional hierarchies due to the presence of topic-oriented OLAP operators. A rather similar approach is proposed in the works of [7] where the authors propose Meta-stars to model topic hierarchies in ROLAP systems. Its basic idea is to use meta-modeling coupled with navigation tables and with traditional dimension tables: navigation tables support hierarchy instances with different lengths and with non-leaf facts, and allow different roll-up semantics to be explicitly annotated; meta-modeling enables hierarchy heterogeneity and dynamics to be accommodated; dimension tables are easily integrated with standard business hierarchies.

Further to this study, we may conclude that most of these works ensure a special treatment of tweets but do not offer tools for the decision-makers to manipulate the information contained in the combined meta-data associated with their tweets.

Hence, we aim at providing a generic multidimensional model supporting tweets i.e., independent of the special needs pre-defined a priori while considering the structural specificity and possibly semantic data.

3 Multidimensional Modeling of Tweets

In this section we introduce the multidimensional model dedicated to the OLAP of tweets. We then suggest some extensions to reflect the specificities of data extracted from tweets (Tweet/Tweet-responses and calculated attributes). The concepts of our multidimensional model are already presented in our previous work [1]. We only present the formalism of extended concepts.

We have extended the conventional concept of fact by adding a ***reflexive relationship***, denoted $\mathcal{R}$, between ***the fact instances*** that allows connecting an instance of the fact to one or several instances of the same fact. This relationship will guarantee that every Tweet response added to the table corresponds to an existing Tweet and then analyses of linked tweets are possible. In addition to that, there exist data extracted directly from tweets (raw data) and calculated data. For this reason, we extend the concept of dimension.

- $\forall$ i $\in$ [1...n], **a fact $\mathbf{F_i}$** is defined by ($NameF_i$; M_i ; INS_i ; $\mathcal{R}_i$) where:
  - $NameF_i$ is the name identifying the fact F_i in the constellation,
  - $M_i = \{m_{i1}, ..., m_{ik}\}$ is a set of k measures of F_i,
  - $INS_i = \{ins_{i1}, ..., ins_{il}\}$ is the set of l instances of fact F_i,
  - $\mathcal{R}_i : INS_i \rightarrow INS_i$, as $\mathcal{R}_i(INS_i) \subseteq INS_i$.
- i $\in$ [1...m], **a dimension D_i** is defined by ($NameD_i$; A_i; A'_i; H_i) where:
  - $NameD_i$ is the name identifying the dimension in the constellation,

As a complement to the first proposal, including basic attributes, we extend the definition of the dimension by adding the *calculated attributes*.

- $A_i = \{a_{i1}, ..., a_{iz}\}$ is the set of dimension attributes (parameters and weak attributes) extracted from *raw data*,
- $A^r_i = \{a'_{i1}, ..., a'_{iq}\}$ is the set of dimension attributes extracted from *calculated data*,
- $H_i = \{h_1, ..., h_{ip}\}$ is the set of p hierarchies showing the arrangement of the attributes of D.

According to their calculation method, we distinguish two types of attributes: *ETL calculated* and *runtime calculated*. Values of ETL calculated attributes are directly computed by the ETL process, i.e., at loading; runtime *calculated* have their values derived at the time of analysis by applying various methods and functions.

ETL calculated data

- A new attribute *Tweet-Type*[1]. This helps classifying tweets into one of the following four types: "***Normal-Tweet"*** (Every message comprising less than 140

[1] https://support.twitter.com/articles/119138-types-of-tweets-and-where-they-appear

characters posted on Twitter); "***Mention"*** (Tweets containing the Twitter username of another user, prefixed by the "@" symbol); **"*Responses*" (**Tweet beginning with the username of another user, and is in response to one of his Tweets); **"*Retweet*"** (A tweet starting with the symbol RT).

- Use of additional data sources and APIs: Inclusion of external sources provides an opportunity to add new elements to the multidimensional model. Here are some examples of detection techniques relevant for enriching the Twitter data:
  - Language detection (*Tweet-Language*). Language detection APIs, such as the one offered by Google or JSON, provide such service. Once detected, the language information can be used for analysis.
  - Sentiment detection (*Tweet–Sentiment*) assesses the overall emotion of the content of a tweet (such as positive, negative or neutral). AlchemyAPI [9] is an example of platforms enabling this type of analysis.
  - Topic detection (*Tweet -Topic*) enriches the model with topic assignment. Twitter's own SearchAPI or AlchemyAPI can be used to retrieve daily trending topics.

The miscellaneous elements (*Tweet-Type, Tweet-Language, Tweet–Sentiment, Tweet–Topic)* are grouped into a single ***Junk Dimension***, named *Tweet-Metadata*. [8] defines a "*Junk dimension as a convenient grouping of flags and attributes to get them out of a fact table into a useful dimensional framework*".

Runtime calculated data

- A new hierarchy: We introduce the additional hierarchy User-Category to the USER dimension based on the user's category in terms of the number of that user's followers and friends. There exist different methods dealing with the classification of the Twittos, but most of them agree on the prevailing role of the number of followers and friends. We define formula (1) in order to classify users into three categories.

$$User\ Category = FollowerCount / FriendCount \quad (1)$$

  These categories are: i) **Information Seeker:** Person who posts rarely (*User Category* < 0.8), but follows other users' regularly, ii) **Information Sharing:** This user posts tweets frequently (*User Category* > 1) and has a large number of followers due to the valuable contents of his tweets, and iii) **Friendship relationship:** Equivalence between friends and subscribers ($0.8 \leq$ *User Category* ≤ 1).

- A new hierarchy *User-Activity:* It is added to the *User* dimension and represents the frequency of tweeting (*StatusCount*) relative to the period elapsed since the creation date of the user's account (*TimeDif*). We note that other studies in the literature have adopted the number of retweet to assess the activity of users [5]. However, with the number of retweet, we lose a part of the user activity (i.e., her/his own tweets). We define formula (2) to determine the user activity as

$$User\ Activity = StatusCount / TimeDif \quad (2)$$

  This category should classify each user into one of the following four clusters: **"Old-Active"**, **"New-Active"**, **"Old-Passive"**, and **"New-Passive"**

respectively for those users who registered long ago or recently and who tweet more or less frequently.

Figure 1 shows the multidimensional model of tweets enriched with these extensions. Graphically, calculated data are represented by dashed lines. The model for tweets presents some specifics. The cardinality 0 of a reflexive fact means that a tweet is not necessarily an answer to another tweet. The second specificity is relative to the possibility of having tweets without any associated locality (absence of the PLACE dimension). This aspect is taken into account by our model. Indeed, we defined a relationship of type 1:0 between the fact Activity-Tweet and the PLACE dimension.

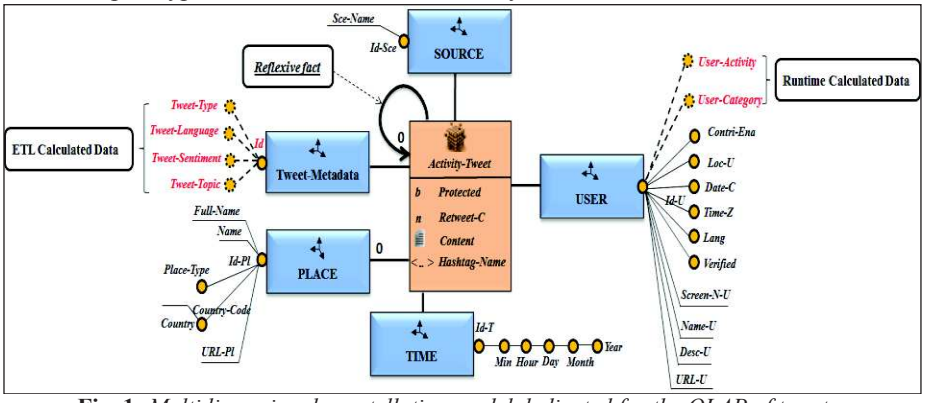

Fig. 1. *Multidimensional constellation model dedicated for the OLAP of tweets.*

4 Experimental results

In order to evaluate our approach we have elaborated a software prototype called *OLAP4Tweets,* developed using JAVA and ORACLE 10g database, since it offers a stable database environment. Fig. 2 gives a functional view of the overall warehousing process of tweets, which is composed of four modules namely:

The ***data source*** is represented by the available Twitter APIs for data streaming. The dataset delivered by the Twitter Streaming API is semi-structured data file conform to JSON (JavaScript Object Notation) output format. Each tweet is streamed as an object containing 67 data fields [10].

The ***multidimensional schema design*** module is three steps based. It aims to propose a multidimensional model dedicated to conventional online analytical processing and, in addition, should allow more elaborate treatments of tweets.

The ***ETL*** module takes care of capturing the original data stream, bringing it into a format compliant with the target database and feeding automatically the various components of the multidimensional model (fact, dimensions, and parameters) issued from the tweets by using Hibernate software and Oracle 10g. In fact, we transform the proposed model into R-OLAP logical model according to the transformation rules for the denormalized R-OLAP model. For a reflexive fact, the primary key contains an additional attribute (Id-Activity-Twt) and a foreign key (Id-Activity-Twt-P) which can contain either a *null value*, or *only values from an existing tweet* (Id-Activity-Twt)

The ***Querying module***. Once the multidimensional model is generated and loaded with data, the decision maker can perform OLAP analyses on tweets using the OLAP tool offered by the implementation platform (e.g., Oracle Discoverer).

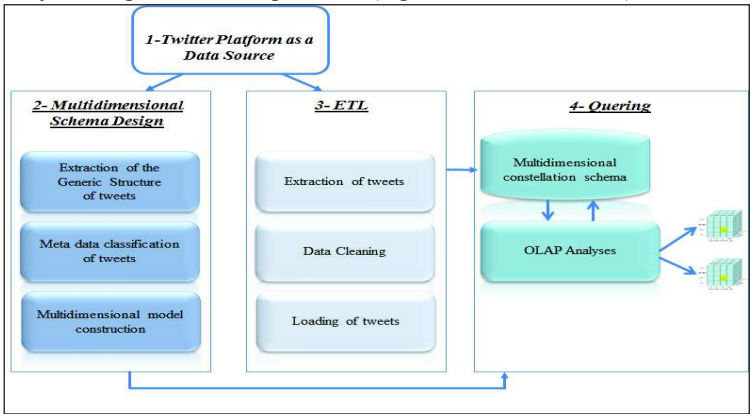

Fig. 2. *Functional view of the overall warehousing process.*

For our experiments, we consider the dataset obtained for the final show of *Arabs Got Talent* (Arab reality television talent show broadcast by MBC4 TV in the Arab world) which was held on March 7, 2015 at 18:00 GMT. We were able to crawl about half a million tweets encompassing (492.826 tweets) 3 hours starting from the beginning of the final show. These tweets are written in different languages. Notice that among these tweets, only 18570 were associated with a place, and 154,800 are response-tweets.

We start by studying the new type of *reflexive fact*. Figure 3 presents sample query using the reflexive fact Activity-Tweet. Table 2 shows the result of this query.

```
SELECT LEVEL, IDACTIV, CONTENTTWEET, SCREEN_N_U,
  IDACTIVREPONSE
FROM   ACTIVITY_TWEET A1, ACTIVITY_TWEET A2, DUSER D
WHERE  A1.Id-Activity-Twt = A2. Id-Activity-Twt-P
AND  A1.ID_U = D.ID_U
CONNECT BY PRIOR IDACTIV = '560673867825031297'
START WITH IDACTIVREPONSE IS NULL
ORDER BY IDACTIV
```

Fig. 3. Example of reflexive OLAP query

Table 1. Result of the query

Level	IDACTIV	CONTENTTWEET	SCREEN-NAME-USER	IDACTIVREPONSE
1	560673867825031297	I would love to see more of this amazing performance! #**MarawaTheAmazing** #ArabsGotTalent	@MahmoudSNasser	NULL
2	560674721235677158	**@MahmoudSNasser** Always	@jamietagg	560673867825031297

		energetic performance		
3	560874826145773641	**@jamietagg** Yes, Beautiful	@Whatweseee	560674721235677158

Conversational posts provide the building blocks of the social interaction between users which leads to the development of community, creation of interpersonal relationships, and the perception of reciprocity between Twitter users and their followers. These conversations are based on *Tweet-Response*. A conversation is intense if it is qualified by more than 5 replies. Table 1 shows all detected intense conversation.

Table 1. Intense conversation

Id-Tweet	Number of Tweet Responses
530401489252937729	5
530428702124146689	7
530775621556002816	9
530391418645516288	6
530460963141844994	27

A Tweet-Content (*Content*) field must contain some content with a maximum length of 140 characters. This lengthy field is fundamental for the semantic analysis as they deliver valuable information about users and their opinions. We have used *AlchemyAPI* to semantically analyze the Tweet-Content. Table 3 shows the distribution of results for the Sentiment Analysis performed on the dataset of our experiment. Fig. 4 depicts sentiments across the top contestants with a variety of talents during the final show.

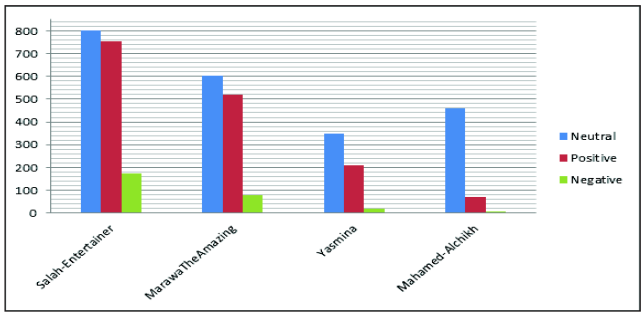

Fig. 4. Sentiment distribution for top contestants tweeted during the final show

Table 2. Sentiment Analysis statistics

Sentiment	Tweet count
Negative	27,858
Neutral	74,247
No Sentiment	324,725
Positive	65,996

5 Conclusion

In this work we have extracted multidimensional data cubes, for OLAP (On-Line Analytical Processing) analyses, from the Twitter social network. We have proposed a multidimensional model dedicated to the OLAP of the data exchanged through tweets

and then we have ensured that this model is generic, that is, not limited to a set of predetermined analytical requirements. Besides, we also have considered the specificities of data tweets: reflexive links between tweets and tweets-response. For that purpose, we have extended the classical concept of multidimensional fact by the proposal of a new type of fact named reflexive fact. We also handled the process of adding new elements such as dimension and hierarchies. Furthermore, our approach was tested on the dataset of the Twitter's public stream with a focus on getting more insight into the content. For this purpose, we presented further examples for sentiment Analysis. We currently continue to perform other OLAP experiments on a larger number of tweets.

Several perspectives for this work are possible. It would be interesting to define new OLAP operators that take into consideration the specificities of this new multidimensional model (Reflexive–Fact). We also expect to exploit the "Text Mining" techniques in order to extract knowledge from tweets and strengthen more semantics in the generic model here proposed.

References

1. Ben Kraiem M., Feki J, Khrouf K, Ravat F, Teste O (2014). OLAP of the Tweets: From Modeling toward exploitation. 8th International Conference on Research Challenges in Information Science (IEEE RCIS'2014), May 28-30, 2014, Marrakesh, Morocco, pp. 45-55.
2. Michelson M and Macskassy S. A. Discovering users' topics of interest on Twitter: a first look. In Proc. AND, pages 73-80, 2010.
3. Mathioudakis, M and Koudas, N. Twittermonitor: trend detection over the twitter stream, in Proceedings of 2010 International Conference on Management of Data, SIGMOD 2010.
4. Bringay, S.,Laurent, A., Poncelet, P., Roche, M., Teisseire, M. Towards an On-Line Analysis of Tweets Processing, 22nd International Conference on Database and Expert Systems Applications, DEXA, Toulouse, France, 2011
5. Rehman, N., Mansmann, S., Weiler, A., Scholl, M.H. Building a Data Warehouse for Twitter Stream Exploration. ACM Fifteenth International Workshop on Data Warehousing and OLAP, DOLAP 2012.
6. Dayal, U., Gupta, C., Castellanos, M., Wang, S., Garcia-Solaco, M. Of cubes, DAGs and hierarchical correlations: A novel conceptual model for analyzing social media data, in: Proceeding. ER, 2012, pp. 30-49.
7. Gallinucci, E., Golfarelli, M., Rizzi, S., Meta-stars: multidimensional modeling for social business intelligence, in: Proceeding of ACM International Workshop on Data Warehousing and OLAP, DOLAP, 2013, pp. 11-18.
8. Kimball, R. The data warehouse toolkit: practical techniques for building dimensional data warehouses, John Wiley & Sons, 1996.
9. AlchemyAPI,Alchemyapi:TransformingTextintoKnowledge http://www.alchemyapi.com⟩, 2008